

CLAIMS

What is claimed is:

- 1 1. A method for performing parallel operations on a pair of objects in a system that
2 includes a plurality of nodes to enable using an increased degree of parallelism, the
3 method comprising the computer-implemented steps of:
4 distributing first-phase partition-pairs of a parallel partition-wise operation on the pair
5 of objects among the plurality of nodes;
6 at a particular node of said plurality of nodes, performing the steps of:
7 partitioning the one or more first-phase partition-pairs distributed to the
8 particular node to produce a set of second-phase partition-pairs; and
9 assigning each second-phase partition-pair from the set of second-phase
10 partition-pairs to a separate slave process.
- 1 2. The method of Claim 1 wherein the step of assigning each second-phase partition-pair
2 from the set of second-phase partition-pairs to a separate slave process is performed
3 by assigning each second-phase partition-pair from the set of second-phase partition-
4 pairs to a separate slave process within said particular node.
- 1 3. The method of Claim 1, wherein the parallel partition-wise operation is a parallel full
2 partition-wise operation.
- 1 4. The method of Claim 1, wherein the parallel partition-wise operation is a parallel
2 partial partition-wise operation.

1 5. The method of Claim 1, wherein the step of partitioning the one or more first-phase
2 partition-pairs and the step of assigning second-phase partition-pairs are performed at
3 each node that has multiple slave processes available for participating in said parallel
4 partition-wise operation.

1 6. The method of Claim 1, further comprising the steps of:
2 determining whether a total number of slave processes available for participating in a
3 second parallel partition-wise operation has a particular logical relationship to
4 a number of first-phase partition-pairs of the second parallel partition-wise
5 operation;
6 if the total number of slave processes available for participating in the second parallel
7 partition-wise operation has said particular logical relationship to the number
8 of first-phase partition-pairs of the second parallel partition-wise operation,
9 then at said particular node performing the steps of:
10 partitioning the one or more first-phase partition-pairs distributed to the
11 particular node for the second parallel partition-wise operation to
12 produce a set of second-phase partition-pairs; and
13 assigning the second-phase partition-pairs from the set of second-phase
14 partition-pairs to slave processes within the particular node to cause the
15 number of slave processes participating in said second parallel
16 partition-wise operation on said particular node to be greater than the
17 number of first-phase partition-pairs that were distributed to said
18 particular node;

19 if the total number of slave processes available for participating in the second
20 parallel partition-wise operation does not have the particular logical
21 relationship to the number of first-phase partition-pairs of the second
22 parallel partition-wise operation, then distributing said first-phase
23 partition-pairs to slave processes without performing second-phase
24 partitioning.

1 7. The method of Claim 6, wherein the step of partitioning the one or more first-phase
2 partition-pairs and the step of assigning the second-phase partition-pairs are
3 performed at each node that has more slave processes available for participating in
4 said second parallel partition-wise operation than the number of first-phase partition-
5 pairs that are distributed to the node.

1 8. The method of Claim 6, wherein the total number of slave processes available for
2 participating in the second parallel partition-wise operation has the particular logical
3 relationship to the number of first-phase partition-pairs of the second parallel
4 partition-wise operation if the total number of slave processes available for
5 participating in the second parallel partition-wise operation is greater than the number
6 of first-phase partition-pairs of the second parallel partition-wise operation.

1 9. The method of Claim 6, wherein the total number of slave processes available for
2 participating in the second parallel partition-wise operation has the particular logical
3 relationship to the number of first-phase partition-pairs of the second parallel
4 partition-wise operation if the total number of slave processes available for

5 participating in the second parallel partition-wise operation is at least an order of
6 magnitude greater than the number of first-phase partition-pairs of the second parallel
7 partition-wise operation.

1 10. The method of Claim 1, wherein the step of distributing first-phase partition pairs is
2 performed based, at least in part, on node affinity with respect to the one or more first-
3 phase partition-pairs of the parallel partition-wise operation and availability of slave
4 processes for performing the parallel partition-wise operation.

1 11. A computer-readable medium carrying instructions for performing parallel operations
2 on a pair of objects in a system that includes a plurality of nodes to enable using an
3 increased degree of parallelism, the instructions comprising instructions for
4 performing the computer-implemented steps of:
5 distributing first-phase partition-pairs of a parallel partition-wise operation on the pair
6 of objects among the plurality of nodes;
7 at a particular node of said plurality of nodes, performing the steps of:
8 partitioning the one or more first-phase partition-pairs distributed to the
9 particular node to produce a set of second-phase partition-pairs; and
10 assigning each second-phase partition-pair from the set of second-phase
11 partition-pairs to a separate slave process.

1 12. The computer-readable medium of Claim 11 wherein the step of assigning each
2 second-phase partition-pair from the set of second-phase partition-pairs to a separate

3 slave process is performed by assigning each second-phase partition-pair from the set
4 of second-phase partition-pairs to a separate slave process within said particular node.

1 13. The computer-readable medium of Claim 11, wherein the parallel partition-wise
2 operation is a parallel full partition-wise operation.

1 14. The computer-readable medium of Claim 11, wherein the parallel partition-wise
2 operation is a parallel partial partition-wise operation.

1 15. The computer-readable medium of Claim 11, wherein the step of partitioning the one
2 or more first-phase partition-pairs and the step of assigning second-phase partition-
3 pairs are performed at each node that has multiple slave processes available for
4 participating in said parallel partition-wise operation.

1 16. The computer-readable medium of Claim 11, further comprising instructions for
2 performing the steps of:
3 determining whether a total number of slave processes available for participating in a
4 second parallel partition-wise operation has a particular logical relationship to
5 a number of first-phase partition-pairs of the second parallel partition-wise
6 operation;
7 if the total number of slave processes available for participating in the second parallel
8 partition-wise operation has said particular logical relationship to the number
9 of first-phase partition-pairs of the second parallel partition-wise operation,
10 then at said particular node performing the steps of:

11 partitioning the one or more first-phase partition-pairs distributed to the
12 particular node for the second parallel partition-wise operation to
13 produce a set of second-phase partition-pairs; and
14 assigning the second-phase partition-pairs from the set of second-phase
15 partition-pairs to slave processes within the particular node to cause the
16 number of slave processes participating in said second parallel
17 partition-wise operation on said particular node to be greater than the
18 number of first-phase partition-pairs that were distributed to said
19 particular node;
20 if the total number of slave processes available for participating in the second
21 parallel partition-wise operation does not have the particular logical
22 relationship to the number of first-phase partition-pairs of the second
23 parallel partition-wise operation, then distributing said first-phase
24 partition-pairs to slave processes without performing second-phase
25 partitioning.

1 17. The computer-readable medium of Claim 16, wherein the step of partitioning the one
2 or more first-phase partition-pairs and the step of assigning the second-phase
3 partition-pairs are performed at each node that has more slave processes available for
4 participating in said second parallel partition-wise operation than the number of first-
5 phase partition-pairs that are distributed to the node.

1 18. The computer-readable medium of Claim 16, wherein the total number of slave
2 processes available for participating in the second parallel partition-wise operation has

the particular logical relationship to the number of first-phase partition-pairs of the second parallel partition-wise operation if the total number of slave processes available for participating in the second parallel partition-wise operation is greater than the number of first-phase partition-pairs of the second parallel partition-wise operation.

19. The computer-readable medium of Claim 16, wherein the total number of slave processes available for participating in the second parallel partition-wise operation has the particular logical relationship to the number of first-phase partition-pairs of the second parallel partition-wise operation if the total number of slave processes available for participating in the second parallel partition-wise operation is at least an order of magnitude greater than the number of first-phase partition-pairs of the second parallel partition-wise operation.

20. The computer-readable medium of Claim 11, wherein the step of distributing first-phase partition pairs is performed based, at least in part, on node affinity with respect to the one or more first-phase partition-pairs of the parallel partition-wise operation and availability of slave processes for performing the parallel partition-wise operation.

21. A method for performing parallel operations on a pair of objects including a source object and a target object in a broadcasting operation, the method comprising the computer-implemented steps of:
mapping each tuple from a source object to a corresponding static partition of a plurality of static partitions of the target object;

6 distributing the static partitions among the plurality of nodes according to a node
7 distribution criteria; and
8 assigning each static partition to a slave process; and
9 broadcasting each tuple only to a group of slave processes assigned to the static
10 partition to which the tuple is mapped.

1 22. The method of Claim 21, wherein the node distribution criteria includes node affinity
2 with respect to the one or more first-phase partition-pairs of the parallel partition-wise
3 operation and availability of slave processes for performing the parallel partition-wise
4 operation.

1 23. A computer-readable medium carrying instructions for performing parallel operations
2 on a pair of objects including a source object and a target object in a broadcasting
3 operation, the instructions comprising instructions for performing the computer-
4 implemented steps of:
5 mapping each tuple from a source object to a corresponding static partition of a
6 plurality of static partitions of the target object;
7 distributing the static partitions among the plurality of nodes according to a node
8 distribution criteria; and
9 assigning each static partition to a slave process; and
10 broadcasting each tuple only to a group of slave processes assigned to the static
11 partition to which the tuple is mapped.

- 1 24. The computer-readable medium of Claim 23, wherein the node distribution criteria
2 includes node affinity with respect to the one or more first-phase partition-pairs of the
3 parallel partition-wise operation and availability of slave processes for performing the
4 parallel partition-wise operation.